

Probability bounds for active learning in the regression problem

Ana K. Fermín

Modal'X, Université Paris Ouest, France
e-mail: aferminrodriguez@u-paris10.fr

and

Carenne Ludeña*

Facultad de Ciencias, Universidad Central de Venezuela
e-mail: carinludena@gmail.com

Abstract: In this article we consider the problem of choosing an optimal sampling scheme for the regression problem simultaneously with that of model selection. We consider a batch type approach and an on-line approach following algorithms recently developed for the classification problem. Our main tools are concentration-type inequalities which allow us to bound the supremum of the deviations of the sampling scheme corrected by an appropriate weight function.

AMS 2000 subject classifications: Primary 68T05, 62J05, 62K05.

1. Introduction

Consider the following regression model

$$y(t_i) = x_0(t_i) + \varepsilon_i, \quad (1)$$

where the observation noise ε_i are i.i.d. realizations of a certain random variable ε . The problem we consider in this article is that of estimating x_0 based on a subsample of size $N \ll n$ of the data collection

$$(t_1, y_1), \dots, (t_n, y_n).$$

This occurs when, for example, obtaining the values of y_i for each sample point t_i is expensive or time consuming or because it is necessary to set up an experimental design based on previous data.

Let \hat{x}_N be the chosen estimator. Intuitively we would like that

$$\|x_0 - \hat{x}_N\| \sim \|x_0 - \hat{x}_n\|,$$

where \hat{x}_n is the “best” possible estimator in some sense over the whole data collection, with N small. That is, a good sample selection requires searching for the most informative, in some sense, part of the sample.

In this article we propose a statistical regularization approach for selecting a good subsample of the data by introducing a weighted sampling scheme (importance weighting) and an appropriate penalty function over the sampling choices. This will be done by fixing a spanning family $\{\phi_j\}_j$ and considering the best approximation x_m of x_0 over $\{\phi_j\}_{j=1}^m$. In this way the problem of model selection and choosing a good sampling set can be considered simultaneously. This is what is

*This work was partially supported by Project CDCH PG-03-8304-2011/1

known as active learning. We will consider two approaches. The first, the batch approach [10], assumes the sampling set is chosen all at once, based on the minimization of a certain penalized loss function which can then be generalized to consider the problem of selecting the model as well. The second, the iterative approach [2], considers a two step iterative method choosing alternatively the best new point to be sampled and the best model given the set of points. In both cases, based on concentration-type inequalities we will show that the estimation schemes attain optimal rates while reducing the size of the sample. Although variance minimization techniques for choosing appropriate subsamples is a well used tool in practice, giving adequate bounds in probability allowing for active learning which leads to optimal rates has been much less studied in the regression setting.

The article is organized as follows. In section 2 we formulate the basic problem and study a batch approach for simultaneous sample and model selection. In section 3 we study an iterative approach to sample selection and we discuss effective sample size reduction. Section 4 is devoted to the proof of the more technical results.

2. Preliminaries

2.1. Formulation of the problem and basic assumptions

We are interested in recovering a certain approximation of x_0 based on observations

$$y_i = x_0(t_i) + \varepsilon_i, \quad i = 1, \dots, n$$

where ε_i are i.i.d. realizations of a random variable ε satisfying the moment condition

MC Assume the r.v. ε satisfies $\mathbb{E}\varepsilon = 0$, $\mathbb{E}(|\varepsilon|^k/\sigma^k) \leq k!/2$ for all $k > 2$ and $\mathbb{E}(\varepsilon^2) = \sigma^2$.

We also need some notation concerning the fixed design, $t_i, i = 1, \dots, n$. Let

$$\delta_{t_i}(t) = \begin{cases} 1, & \text{if } t = t_i; \\ 0, & \text{if not.} \end{cases}$$

Define the empirical measure:

$$P_n = \frac{1}{n} \sum_{i=1}^n \delta_{t_i},$$

the associated empirical norm

$$\|y\|_n^2 = \|y\|_{P_n}^2 = \frac{1}{n} \sum_{i=1}^n (y(t_i))^2,$$

and the empirical scalar product

$$\langle y, u \rangle_n = \frac{1}{n} \sum_{i=1}^n u(t_i)y(t_i).$$

With the above notation, given any positive function r , we also introduce the r -scalar product $\langle y, u \rangle_{n,r} = \frac{1}{n} \sum_{i=1}^n r(t_i)u(t_i)y(t_i)$ and $\|y\|_{n,r}$ the associated empirical norm.

2.2. Discretization scheme

To start with we will consider the approximation of x_0 over a finite-dimensional subspace S_m . This subspace will be assumed to be linearly spanned by the set $\{\phi_j\}_{j \in I_m} \subset \{\phi_j\}_{j \geq 1}$, with I_m a certain index set.

We assume there exists a certain density q such that

AQ There exists a positive constant Q such that $q(t_i) \leq Q, i = 1, \dots, n$ and

$$\int \phi_l(u) \phi_k(u) q(u) du = \delta_{k-l}.$$

We will also require the following assumption.

AB There exists an increasing sequence c_m such that $\|\phi_j\|_\infty \leq c_m$ for $j \leq m$.

Let $G_m = [\phi_j(t_i)]_{i,j}$ be the associated empirical $n \times m$ Gram matrix (design matrix), so that $\frac{1}{n} G_m^t D_q G_m \rightarrow I_m$, where D_q is the diagonal matrix with entries $q(t_i)$, for $i = 1, \dots, n$.

We will assume the following approximation property for S_m

AS There exist positive constants α and $c_1 < c_2$, such that

$$c_1 n^{-1-\alpha} \leq \|I_m - \frac{1}{n} G_m^t D_q G_m\|_\rho \leq c_2 n^{-1-\alpha},$$

where for any matrix A , $\|A\|_\rho$ stands for the usual spectral norm of A in the L_2 norm.

We will denote by $\hat{x}_m \in S_m$ the function that minimizes the weighted norm $\|x - y\|_{n,q}^2$ over S_m . This is,

$$\hat{x}_m = \arg \min_{x \in S_m} \frac{1}{n} \sum_{i=1}^n (y_i - x(t_i))^2 q(t_i) = R_m y,$$

with $R_m = G_m (G_m^t D_q G_m)^{-1} G_m^t D_q$ the orthogonal projector over S_m in the q -empirical norm $\|\cdot\|_{n,q}$.

Let $x_m := R_m x_0$ be the projection of x_0 over S_m in the q -empirical norm $\|\cdot\|_{n,q}$. Our goal is to choose a good subsample of the data collection such that the estimator of the unobservable function x_0 in the finite-dimensional subspace S_m , based on this subsample, attains optimal error bounds. For this we must introduce the notion of subsampling scheme and importance weighted approaches (see [2], [10]), which we discuss below.

2.3. Sampling scheme and importance weighting algorithm

In order to sample the data set we will introduce a sampling probability $p(t)$ and a sequence of Bernoulli($p(t_i)$) random variables w_i , $i = 1, \dots, n$ independent of ε_i with $p(t_i) > p_{\min}$. Let $D_{w,q,p}$ be the diagonal matrix with entries $q(t_i)w_i/p(t_i)$. So that $\mathbb{E}(D_{w,q,p}) = D_q$. Sometimes it will be more convenient to rewrite $w_i = 1_{u_i < p_i}$ for $\{u_i\}_i$ and i.i.d. sample of uniform random variables, independent of $\{\varepsilon_i\}_i$ in order to stress the dependence on p of the random variables w_i .

The next step is to construct an estimator for $x_m = R_m x_0$, based on the observation vector y and the sampling scheme p . For this, we consider a modified version of the estimator $\hat{x}_m = R_m y$.

As the approximation of x_m , we then take (for a fixed m and p)

$$\begin{aligned} \hat{x}_{m,p} &= \arg \min_{x \in S_m} \|x - y\|_{n, \frac{q w}{p}}^2 \\ &= \arg \min_{x \in S_m} \frac{1}{n} \sum_{i=1}^n (y_i - x(t_i))^2 \frac{w(t_i)}{p(t_i)} q(t_i). \end{aligned}$$

that is,

$$\hat{x}_{m,p} = R_{m,p}y, \quad (2)$$

for $R_{m,p} = G_m(G_m^t D_{w,q,p} G_m)^{-1} G_m^t D_{w,q,p}$ the orthogonal projector over S_m in the wq/p -empirical norm $\|\cdot\|_{n,wq/p}$. Note that this estimator depends on y_i only if $w(t_i) = 1$.

2.4. Choosing a good sampling scheme (for a fixed model)

Here, we will assume that S_m is fixed with dimension $|I_m| = m$. In this case we will assume that the bias $\|x_0 - x_m\|_{n,q}^2$, which is independent of p , is known up to a constant (for example based on approximation errors over a fixed model space) and study rather the approximation error $\|x_m - \hat{x}_{m,p}\|_{n,q}^2$. The latter depends on how the sampling probability p is chosen.

Let $\mathcal{P} := \{P_k : \{t_1, \dots, t_n\} \rightarrow [0, 1]^n, k \geq 1\}$ be a numerable collection of $[0, 1]$ valued functions. We will assume that $\min_k \min_i P_k(t_i) > p_{\min}$.

A good sampling scheme p , based on the data, should be the minimizer of the non observable quantity $\|x_m - \hat{x}_{m,p}\|_{n,q}^2$. To overcome this difficulty, we observe that since $R_{m,p}x_m = x_m$, then

$$\begin{aligned} & [x_m - \hat{x}_{m,p}] \\ &= R_{m,p}[x_0 - x_m] + R_{m,p}\varepsilon \\ &= \mathbb{E}(R_{m,p})[x_0 - x_m] + (R_{m,p} - \mathbb{E}(R_{m,p}))[x_0 - x_m] + R_{m,p}\varepsilon. \end{aligned} \quad (3)$$

Consider the deterministic term $\mathbb{E}(R_{m,p})[x_0 - x_m]$. We shall prove in Lemma 2.4.5 that under condition [AS], the term $\|\mathbb{E}(R_{m,p})[x_0 - x_m]\|_{n,q}$ is of order $O(n^{-1-\alpha}\|x_0 - x_m\|_{n,q}/p_{\min})$. Whence, any minimizer should essentially account for the biggest possible values, with high probability, of the second and third terms. It is thus reasonable, to consider the best p as the minimizer

$$\hat{p} = \underset{p \in \mathcal{P}}{\operatorname{argmin}} \widetilde{pen}(m, p, \delta, \gamma, n), \quad (4)$$

where, for a given $0 < \gamma < 1$,

$$\widetilde{pen}(m, p, \delta, \gamma, n) = \{(1 + \gamma)\widetilde{pen}_1(m, p, \delta) + (1 + 1/\gamma)\widetilde{pen}_2(m, p, \delta)\}$$

with pen_1 and pen_2 , which will be defined below, such that

$$\begin{aligned} & P(\sup_{\mathcal{P}} \{\|(R_{m,p} - \mathbb{E}(R_{m,p}))[x_0 - x_m]\|_{n,q}^2 - \widetilde{pen}_1(m, p, \delta)\} > 0) < \delta/2, \\ & P(\sup_{\mathcal{P}} \{\|R_{m,p}\varepsilon\|_{n,q}^2 - \widetilde{pen}_2(m, p, \delta)\} > 0) < \delta/2. \end{aligned}$$

The last two inequalities will be examined separately in Lemma 2.4.1 and Lemma 2.4.3. These Lemmas together with Lemma 2.4.5 and the definition of the penalization terms assure that the proposed estimation procedure is not only consistent but that it achieves optimal rates.

For each $p \in \mathcal{P}$, let $k(p)$ be its corresponding index and define

$$\widetilde{pen}_1(m, p, \delta) = \|x_0 - x_m\|_{n,q}^2 (\tilde{\beta}_{m,k(p)} (1 + \tilde{\beta}_{m,k(p)}^{1/2}))^2 \quad (5)$$

with

$$\tilde{\beta}_{m,k(p)} = \frac{c_m(\sqrt{17} + 1)}{2} \sqrt{\frac{mQ}{np_{\min}}} \sqrt{2 \log(2^{7/4} m k(p)(k(p) + 1)/\delta)}, \quad (6)$$

and finally,

$$\begin{aligned} \widetilde{pen}_2(m, p, \delta) &= \sigma^2 r Q (1 + L_{k(p)}) \frac{m+1}{n} \\ &\quad + \sigma^2 Q \frac{\log^2(2/\delta)}{dn}, \end{aligned} \quad (7)$$

with $r > 1$, $d = d(r)$ a positive constant that depends on r and $L_k = L_{k(p)} \geq 0$ a sequence such that the following Kraft condition $\sum_k e^{-\sqrt{dr} L_k (m+1)} < 1$ holds.

We have the following result.

Lemma 2.4.1. *Assume that the conditions $[AB]$, $[AS]$, and $[AQ]$ are satisfied and that there is a constant $p_{min} > 0$ such that for all $i = 1, \dots, n$, $p(t_i) > p_{min}$. Assume \widetilde{pen}_1 to be selected according to (5). Then for all $\delta > 0$ we have*

$$P \left[\sup_{\mathcal{P}} \{ \| (R_{m,p} - \mathbb{E}(R_{m,p})) [x_0 - x_m] \|_{n,q}^2 - \widetilde{pen}_1(m, p, \delta) \} > 0 \right] \leq \delta/2$$

Proof. We will achieve the proof by bounding

$$\| (R_{m,p} - \mathbb{E}(R_{m,p})) [x_0 - x_m] \|_{n,q}^2 \leq \| R_{m,p} - \mathbb{E}(R_{m,p}) \|_{\rho}^2 \| x_0 - x_m \|_{n,q}^2.$$

For this we shall consider a double application of a straightforward generalization of Theorem 7.3 in [8], whose proof is given in the Appendix.

Lemma 2.4.2. *Let $A \in \mathbb{R}^{n \times m}$ be some matrix whose rows, $a(l) \in \mathbb{R}^m$, $l=1, \dots, n$, satisfy $\|a(l)\|_2 \leq K\sqrt{m}$ for some constant $K \geq 1$. Consider the matrix $A^T A = \sum_{l=1}^n a(l)a(l)^t$ and let $\Lambda_A = \frac{1}{n} \mathbb{E}(A^T A) \|_{\rho}$. Set $\tau = (\sqrt{17} + 1)/4$. We have the following bounds:*

- Define $E_r := \mathbb{E} \left(\left\| \frac{1}{n\Lambda_A} (A^T A - \mathbb{E}(A^T A)) \right\|_{\rho}^r \right)$ and let

$$\Sigma_{r,m,n} = \left(\frac{2K}{\sqrt{n\Lambda_A}} \sqrt{\frac{m}{n}} \right)^r 2^{3/4} m r^{r/2} e^{-r/2}.$$

Then for any $r \geq 2$,

$$E_r^{1/r} \leq \tau \Sigma_{r,m,n}.$$

- Let $\delta < 1/2$, then the following bound in probability holds true for $u \geq \sqrt{2}$

$$P \left(\left\| \frac{A^T A - \mathbb{E}(A^T A)}{n\Lambda_A} \right\|_{\rho} > \frac{2\tau K}{\sqrt{\Lambda_A}} \sqrt{\frac{m}{n}} u \right) \leq m 2^{3/4} e^{-u^2/2},$$

or equivalently with probability at least $1 - \delta$

$$\left\| \frac{A^T A - \mathbb{E}(A^T A)}{n\Lambda_A} \right\|_{\rho} \leq \frac{2\tau K}{\sqrt{\Lambda_A}} \sqrt{\frac{m}{n}} \sqrt{2 \log(2^{3/4} m / \delta)}.$$

With this lemma we continue the proof of Lemma 2.4.1. Recall that $R_{m,p} = \frac{1}{n} G_m (\frac{1}{n} G_m^t D_{w,q,p} G_m)^{-1} G_m^t D_{w,q,p}$. On the other hand, observe that since $A_{m,p} := 1/n G_m^T D_{pqw} G_m$ is a positive definite matrix its inverse exists and moreover we may write $A_{m,p}^{-1/2}$ using the standard spectral notation. Also since $A_{m,p}$ is symmetric we have $A_{m,p}^{-1/2} = (A_{m,p}^{-1/2})^t$.

Consider the matrix $\tilde{A}_{m,p} = D_{pqw}^{1/2} G_m$. Then we have its rows $\tilde{a}_{m,p}(l)$ satisfy

$$\|\tilde{a}_{m,p}(l)\|_2 = \sqrt{\frac{w_i q_i}{p_i} \sum_{j=1}^m (\phi_j(t_l))^2} \leq c_m \sqrt{m \frac{Q}{p_{\min}}}$$

and $\Lambda_{\tilde{A}_{m,p}} = \|\mathbb{E} \left(\frac{1}{n} \tilde{A}_{m,p}^T \tilde{A}_{m,p} \right)\|_\rho \leq 1 + c_2(n)^{-\alpha-1}$ under assumption [AS].

In what follows set $k = k(p)$ and let $\delta'_k = \delta/(2k(k+1))$. A first application of Lemma 2.4.2 then yields

$$\|A_{m,p} - \mathbb{E}(A_{m,p})\|_\rho \leq 2\tau c_m \sqrt{1 + c_2(n)^{-\alpha-1}} \sqrt{\frac{mQ}{np_{\min}}} \sqrt{2 \log(2^{3/4} m / \delta'_k)}$$

with probability greater $1 - \delta'_k$. Here, the choice of δ'_k is required in order to account for the supremum over the collection of possible sampling schemes.

It then follows using a classical Neumann series expansion that with probability greater than $1 - \delta'_k$,

$$\|A_{m,p}^{-1}\|_\rho \leq \frac{1}{1 - (\sqrt{1 + c_2 n^{-\alpha-1}} \tilde{\beta}_{m,k} + c_2 n^{-\alpha-1})}, \quad (8)$$

where

$$\tilde{\beta}_{m,k} = 2\tau c_m \sqrt{\frac{mQ}{np_{\min}}} \sqrt{2 \log(2^{3/4} m / \delta'_k)}.$$

Now, consider the matrix $\tilde{E}_{m,p} = A_{m,p}^{-1/2} G_m^t D_{pqw}^{1/2}$ and note that the projection matrix $R_{m,p} = \frac{1}{n} G_m A_{m,p}^{-1} G_m^T D_{pqw} = \frac{1}{n} \tilde{E}_{m,p}^t \tilde{E}_{m,p}$. Using the singular value decomposition and the definition of $\tilde{E}_{m,p}$, we have

$$\|\tilde{E}_{m,p}^T \tilde{E}_{m,p} - \mathbb{E}(\tilde{E}_{m,p}^T \tilde{E}_{m,p})\|_\rho = \|\tilde{E}_{m,p} \tilde{E}_{m,p}^T - \mathbb{E}(\tilde{E}_{m,p} \tilde{E}_{m,p}^T)\|_\rho,$$

since the singular values are the same. Thus, it is enough for our purposes to bound $\|\tilde{E}_{m,p} \tilde{E}_{m,p}^T - \mathbb{E}(\tilde{E}_{m,p} \tilde{E}_{m,p}^T)\|_\rho$ in probability.

Next, we bound the rows of matrix $\tilde{E}_{m,p}^t, \tilde{e}_{m,p}^t(l)$. As before and using the bound in (8) we have

$$\|\tilde{e}_{m,p}^t(l)\|_2 \leq c_m \sqrt{\frac{mQ}{p_{\min}}} [1 + (\sqrt{1 + c_2 n^{-\alpha-1}} \tilde{\beta}_{m,k} + c_2 n^{-\alpha-1})]^{1/2}.$$

On the other hand, because $R_{m,p}$ is a projection matrix, $\|R_{m,p}\|_\rho = 1$ and we have

$$\begin{aligned} 1 &= \mathbb{E}(\|R_{m,p}\|_\rho) \geq \sup_{\|u\|_2=1} \mathbb{E}(\|R_{m,p} u\|_2) \\ &\geq \sup_{\|u\|_2=1} \|\mathbb{E}(R_{m,p}) u\|_2 = \|\mathbb{E}(R_{m,p})\|_\rho \end{aligned}$$

so that $\Lambda_{\tilde{E}_{m,p}^t} \leq 1$.

Then Lemma 2.4.2 yields the stated result by the choice of the penalization $\widehat{pen}_1(m, p, \delta)$ and taking a union bound over $p \in \mathcal{P}$. □

Lemma 2.4.3. Assume the observation noise in equation (1) is an i.i.d. collection of random variables satisfying the moment condition [MC]. Assume that the condition [AQ] is satisfied and assume that there is a constant $p_{\min} > 0$ such that $p(t_i) > p_{\min}$ for all $i = 1, \dots, n$. Assume \widetilde{pen}_2 to be selected according to (7) with $r > 1$, $d = d(r)$ and $L_k \geq 0$, such that the following Kraft inequality $\sum_k e^{-\sqrt{drL_k(m+1)}} < 1$ holds. Then,

$$P(\sup_{\mathcal{P}} \{\|R_{m,p}\varepsilon\|_{n,q}^2 - \widetilde{pen}_2(m, p, \delta)\} > 0) < \delta/2.$$

Proof. For a given positive function f , recall $\|u\|_{f,n}^2 = 1/n \sum_{i=1}^n f(t_i)u(t_i)$. Let $u \in S_m$ and consider a linear application $A : R^n \rightarrow R^m$. Define $\eta(A, z) := \sqrt{z^t A^t A z}$ for $z \in R^n$ and $\eta_f(A, z) := \sup_{\|u\|_{f,n}=1} \sum_{i=1}^n f(t_i)z_i(A^t u)_i$. Then, $\eta_f(A, z) = \eta(AD_f^{1/2}, z)$, where D_f is the diagonal matrix with entries f_i .

On the other hand, note that $\|R_{m,p}\varepsilon\|_{n,q} = \eta(R_{m,p}D_q^{1/2}, \varepsilon)$. The proof then follows directly from the next lemma, whose proof is contained in [7].

Lemma 2.4.4. Let $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^t$ be a vector of i.i.d. random variables satisfying the moment condition [MC]. Let A be a given $m \times n$ matrix. Define $\eta(A) = \eta(A, \varepsilon) = \sqrt{\varepsilon^t A^t A \varepsilon}$. Then, for $r > 1$, $u > 0$ and $L > 0$ there exists a positive constant d that depends on r such that the following inequality holds

$$\begin{aligned} P(\eta^2(A) \geq \sigma^2[Tr(A^t A) + \rho(A^t A)]r(1+L) + \sigma^2 u) \\ \leq \exp\{-\sqrt{d(1/\rho(A^t A)u + rL[Tr(A^t A)/\rho(A^t A) + 1])}\}. \end{aligned}$$

To apply Lemma 2.4.4, we have study the terms of the trace and spectral radius of the matrix $\Gamma = (R_{m,p}D_q^{1/2})^T R_{m,p}D_q^{1/2}$. But, as $R_{m,p}$ is a projection operator then $Tr(\Gamma) \leq Qm$ and the spectral radius $\rho(\Gamma) \leq Q$.

Thus, we have

$$\begin{aligned} & P\left(\sup_{\mathcal{P}} \{\|R_{m,p}\varepsilon\|_{n,q}^2 - \widetilde{pen}_2(m, p, \delta)\} > 0\right) \\ & \leq \delta/2 \times \sum_k \exp\{-\sqrt{drL_k[m+1]}\} \end{aligned}$$

which yields the desired result. \square

The next lemma control the bias term.

Lemma 2.4.5. Under condition [AS] if $m = o(n)$ and $p_{\min}^{-1} = o(n)$, then

$$\|\mathbb{E}(R_{m,p})[x_0 - x_m]\|_{n,q} = O\left(\frac{n^{-1-\alpha}\|x_0 - x_m\|_{n,q}}{p_{\min}}\right).$$

Proof. Recall from Lemma 2.4.1 $A_{m,p} = 1/n G_m^t D_{w,q,p} G_m$ and set $A_m = \mathbb{E}(A_{m,p}) = 1/n G_m^t D_q G_m$. Then $R_{m,p} = 1/n G_m A_{m,p}^{-1} G_m^t D_{w,q,p}$ and

$$R_m = \mathbb{E}(1/n G_m A_m^{-1} G_m^t D_{w,q,p}) = 1/n G_m A_m^{-1} G_m^t D_q.$$

Remark that under condition [AS], $\|A_m - I\|_{\rho} \leq c_2 n^{-1-\alpha}$.

Set $Q_{m,p} = A_{m,p} - I$, so that $\|\mathbb{E}(Q_{m,p})\|_\rho \leq c_2 n^{-1-\alpha}$. Set $K_A := c_m \sqrt{Q/p_{\min}}$. By Lemma 2.4.2 we have, for any $r \geq 2$,

$$\begin{aligned} E_{m,p,r}^{1/r} &:= [\mathbb{E}(\|A_{m,p} - \mathbb{E}(A_{m,p})\|)^r]^{1/r} \\ &\leq \tau \left(\frac{2K_A}{\sqrt{n\Lambda}} \sqrt{\frac{m}{n}} \right)^r \Lambda^{3/4} m r^{r/2} e^{-r/2}. \end{aligned}$$

where $\Lambda = \|\mathbb{E}(A_{m,p})\|_\rho = O(1)$.

Whence, for big enough r , since $m = o(n)$ and $p_{\min}^{-1} = o(n)$, we have $E_{m,p,r}^{1/r} = O(n^{-1-\alpha})$ and thus by Hölder's inequality that $\mathbb{E}(\|Q_{m,p}\|_\rho) \leq O(n^{-1-\alpha})$. The latter yields that, in particular, $\mathbb{E}(\|Q_{m,p}\|_\rho) < 1$. Set $C_{m,p} := A_{m,p}^{-1} - I$. Using the classical Neumann expansion, under condition [AS], by the Monotone Converge Theorem we may finally bound $\mathbb{E}(\|C_{m,p}\|_\rho) \leq c n^{-1-\alpha}$ for a certain positive constant c . We also have $A_m^{-1} = I + C_m$ with $\|C_m\|_\rho \leq n^{-1-\alpha}$.

On the other hand remark, from the definition of the spectral norm, that for any matrix B , $\|B\|_\rho = \|B^T\|_\rho = \sqrt{\|B * B^T\|_\rho}$, so that for any given matrix M ,

$$\begin{aligned} \|1/n G_m M G_m^t D_{w,q,p} [x_0 - x_m]\|_{n,q} &\leq \|M\|_\rho \|1/n G_m G_m^t\|_\rho \|D_{w,q,p}\|_\rho \|x_0 - x_m\|_{n,q} \\ &\leq \|M\|_\rho \|x_0 - x_m\|_{n,q} / p_{\min}, \end{aligned}$$

where the last bound follows from the definition of the diagonal matrix $D_{w,q,p}$ and the bounds on $\|1/n G_m G_m^t\|_\rho$ under condition [AS].

Then, since by definition $R_m[x_0 - x_m] = 0$ we have

$$\begin{aligned} &\|\mathbb{E}(R_{m,p})[x_0 - x_m]\|_{n,q} \\ &= \|\mathbb{E}(1/n G_m [I + C_{m,p} + A_m^{-1} - A_m^{-1}] G_m^t D_{w,q,p}) [x_0 - x_m]\|_{n,q} \\ &\leq \|\mathbb{E}(1/n G_m C_{m,p} G_m^t D_{w,q,p}) [x_0 - x_m]\|_{n,q} \\ &\quad + \|1/n G_m [I - A_m^{-1}] G_m^t \mathbb{E}(D_{w,q,p}) [x_0 - x_m]\|_{n,q} \\ &\leq \mathbb{E}(\|C_{m,p}\|_\rho \|1/n G_m G_m^t\|_\rho \|D_{w,q,p}\|_\rho) \|x_0 - x_m\|_{n,q} \\ &\quad + \|I - A_m^{-1}\|_\rho \|1/n G_m G_m^t\|_\rho \|\mathbb{E}(D_{w,q,p})\|_\rho \|x_0 - x_m\|_{n,q} \\ &\leq c \left[\frac{n^{-1-\alpha}}{p_{\min}} + n^{-1-\alpha} \right] \|x_0 - x_m\|_{n,q}, \end{aligned}$$

where, for the line before last we have used $\|\mathbb{E}(B)\|_\rho \leq \mathbb{E}(\|B\|_\rho)$ for any given matrix B , and the last line follows from the above discussion. \square

Lemmas 2.4.1 and 2.4.3 yield the following consistency result.

Theorem 2.4.6. *Assume that the conditions [AB], [AS] and [AQ] are satisfied and that we use a sampling strategy $p(t_i)$ satisfying*

$$\hat{p} = \operatorname{argmin}_{p \in \mathcal{P}} \{ (1 + \gamma) \widetilde{pen}_1(m, p, \delta) + (1 + 1/\gamma) \widetilde{pen}_2(m, p, \delta) \}$$

with \widetilde{pen}_1 and \widetilde{pen}_2 defined in (5) and (7) for $0 < \gamma < 1$. Then the following inequality holds with probability greater than $1 - \delta$

$$\begin{aligned} &\|x_m - \hat{x}_{m,\hat{p}}\|_{n,q}^2 \\ &\leq 6(\|\mathbb{E}(R_{m,p})(x_m - x_0)\|_{n,q}^2 \\ &\quad + (1 + \gamma) \widetilde{pen}_1(m, p, \delta) + (1 + 1/\gamma) \widetilde{pen}_2(m, p, \delta)). \end{aligned}$$

Proof. For any $p \in \mathcal{P}$,

$$\begin{aligned} & \|x_m - \hat{x}_{m,\hat{p}}\|_{n,q}^2 \\ & \leq \|x_m - \hat{x}_{m,p}\|_{n,q}^2 + \{ \|x_m - \hat{x}_{m,\hat{p}}\|_{n,q}^2 - \widetilde{pen}(m, \hat{p}, \delta, \gamma, n) \} \\ & \quad + \{ \widetilde{pen}(m, p, \delta, \gamma, n) - \|x_m - \hat{x}_{m,p}\|_{n,q}^2 \} \end{aligned}$$

where $\widetilde{pen}(m, p, \delta, \gamma, n) = (1+\gamma)\widetilde{pen}_1(m, p, \delta) + (1+1/\gamma)\widetilde{pen}_2(m, p, \delta)$ with \widetilde{pen}_1 and \widetilde{pen}_2 defined in (5) and (7).

On the other hand, recall that (see equation (3)),

$$x_m - \hat{x}_{m,p} = \mathbb{E}(R_{m,p})[x_0 - x_m] + (R_{m,p} - \mathbb{E}(R_{m,p}))[x_0 - x_m] + R_{m,p}\varepsilon.$$

Since for any $0 < \gamma < 1$, $2ab \leq \gamma a^2 + 1/\gamma b^2$ holds for all $a, b \in \mathbb{R}$, following standard arguments we have

$$\begin{aligned} & \|x_m - \hat{x}_{m,p}\|_{n,q}^2 \\ & \leq 2\|\mathbb{E}(R_{m,p})[x_0 - x_m]\|^2 + 2(1+\gamma)\|(R_{m,p} - \mathbb{E}(R_{m,p}))[x_0 - x_m]\|^2 \\ & \quad + 2(1+1/\gamma)\|R_{m,p}\varepsilon\|_{n,q}^2. \end{aligned}$$

Thus,

$$\begin{aligned} & \|x_m - \hat{x}_{m,\hat{p}}\|_{n,q}^2 \\ & \leq 6\|\mathbb{E}(R_{m,p})(x_m - x_0)\|_{n,q}^2 \\ & \quad + 6(1+\gamma)\widetilde{pen}_1(m, \hat{p}, \delta) + 6(1+1/\gamma)\widetilde{pen}_2(m, \hat{p}, \delta) \\ & \quad + 6(1+\gamma)(\sup_{\mathcal{P}} \{ \|R_{m,p}(x_m - x_0) - \mathbb{E}(R_{m,p})(x_m - x_0)\|_{n,q}^2 \\ & \quad - \widetilde{pen}_1(m, p, \delta) \}) \\ & \quad + 6(1+\gamma^{-1})(\sup_{\mathcal{P}} \{ \|R_{m,p}\varepsilon\|_{n,q}^2 - \widetilde{pen}_2(m, p, \delta) \}) \end{aligned}$$

Finally, as follows from Lemma 2.4.1 and 2.4.3, with probability greater than $1 - \delta$, we have the stated result. \square

2.5. Model selection and active learning

Given a model and n observations $\{x_i, y_i\}_{i=1}^n$ we know how to estimate the best sampling scheme \hat{p} and to obtain the estimator $\hat{x}_{m,\hat{p}}$. The problem is that the model m might not be a good one. Instead of just looking at *fixed* m we would like to consider simultaneous model selection as in [10]. For this we shall pursue a more global approach based on loss functions.

We start by introducing some notation. Set $l(u, v) = (u - v)^2$ the squared loss and let $L_n(x, y, p) = \frac{1}{n} \sum_{i=1}^n q_i \frac{w_i}{p_i} l(x(t_i), y_i)$ be the empirical loss function for the quadratic difference with the given sampling distribution. Set $L(x) := \mathbb{E}(L_n(x, y, p))$ with the expectation taken over all the random variables involved. Let $L_n(x, p) := \mathbb{E}_\varepsilon(L_n(y, x, p))$ where $\mathbb{E}_\varepsilon()$ stands for the conditional expectation given the sample w , that is the expectation with respect to the random noise. It is not hard to see that

$$L(x) = \frac{1}{n} \sum_{i=1}^n q_i \mathbb{E}(l(x(t_i), y_i)),$$

and

$$L_n(x, p) = \frac{1}{n} \sum_{i=1}^n q_i \frac{w_i}{p_i} \mathbb{E}(l(x(t_i), y_i)).$$

Recall that $\hat{x}_{m,p} = R_{m,p}y$ is the minimizer of $L_n(x, y, p)$ over each S_m for given p and that $x_m = R_m x_0$ is the minimizer of $L(x)$ over S_m . Our problem is then to find the best approximation of the target x_0 over the function space $S_0 := \bigcup_{m \in I} S_m$. In the notation of section 2.2 we assume for each m that S_m is a bounded subset of the linearly spanned space of the collection $\{\phi_j\}_{j \in I_m}$ with $|I_m| = d_m$.

Unlike the fixed m setting, model selection requires controlling not only the variance term $\|x_m - \hat{x}_{m,p}\|_{n,q}$ but also the unobservable bias term $\|x_0 - x_m\|_{n,q}^2$ for each possible model S_m . If all samples were available this would be readily available just by looking at $L_n(x, y, p)$ for all S_m and p , but in the active learning setting labels are expensive.

Set $e_m := \|x_0 - x_m\|_\infty$. In what follows we will assume that there exists a positive constant C such that $\sup_m e_m \leq C$. Remark this implies $\sup_m \|x_0 - x_m\|_{n,q} \leq QC$, with Q defined in [AQ].

Recalling $P_k \in \mathcal{P}$ stands for the set of candidate sampling probabilities, set $p_{k,\min} = \min_i(P_{k,i})$. Define

$$pen_0(m, P_k, \delta) = \frac{QC^2}{p_{k,\min}} \sqrt{\frac{1}{2n} \ln\left(\frac{6d_m(d_m+1)}{\delta}\right)}, \quad (9)$$

$$pen_1(m, P_k, \delta) = QC\beta_{m,k}^2(1 + \beta_{m,k}^{1/2})^2, \quad (10)$$

with

$$\beta_{m,k} = \frac{c_m(\sqrt{17}+1)}{2} \sqrt{\frac{d_m Q}{np_{k,\min}}} \sqrt{2 \log\left(\frac{3 * 2^{7/4} d_m^2 (d_m+1) k(k+1)}{\delta}\right)},$$

and finally

$$pen_2(m, P_k, \delta) = Q\sigma^2 \left\{ r(1 + L_{m,k}) \frac{d_m+1}{n} + \frac{\ln^2(6/\delta)}{dn} \right\} \quad (11)$$

where $L_{m,k} \geq 0$ is a sequence such that $\sum_{m,k} e^{-\sqrt{dr L_{m,k}(d_m+1)}} < 1$ holds. We remark that the change from δ to $\delta/(d_m(d_m+1))$ in pen_0 and pen_1 is required in order to account for the supremum over the collection of possible model spaces S_m .

Also, we remark that introducing simultaneous model and sample selection results in the inclusion of term $pen_0 \sim C^2/p_{k,\min} * \sqrt{1/n}$ which includes an L_∞ type bound instead of an L_2 type norm which may yield non optimal bounds. Dealing more efficiently with this term would require knowing the (unobservable) bias term $\|x_0 - x_m\|_{n,q}$. A reasonable strategy is selecting $p_{k,\min} = p_{k,\min}(m) \geq \|x_0 - x_m\|_{n,q}$ whenever this information is available. In practice, $p_{k,\min}$ can be estimated for each model m using a previously estimated empirical error over a subsample if this is possible. However this yields a conservative choice of the bound. One way to avoid this inconvenience is to consider iterative procedures, which update on the unobservable bias term. This course of action shall be pursued in section 3.

With these definitions, for a given $0 < \gamma < 1$ set

$$\begin{aligned} pen(m, p, \delta, \gamma, n) &= [2p_0(m, p, \delta) + \left(\frac{1}{p_{\min}} + \frac{1}{\gamma}\right) pen_1(m, p, \delta) \\ &+ \left(\frac{1}{p_{\min}^2} \left(\frac{2}{\gamma} + 1\right) + \frac{1}{\gamma}\right) pen_2(m, p, \delta) + 2((c+1) \frac{n^{-(1+\alpha)} QC}{p_{\min}})^2]. \end{aligned}$$

and define

$$L_{n,1}(x, y, p) = L_n(x, y, p) + \text{pen}(m, p, \delta, \gamma, n).$$

The appropriate choice of an optimal sampling scheme simultaneously with that of model selection is a difficult problem. We would like to choose simultaneously m and p , based on the data in such a way that optimal rates are maintained. We propose for this a penalized version of $\hat{x}_{m,\hat{p}}$, defined as follows.

We start by choosing, for each m , the best sampling scheme

$$\hat{p}(m) = \arg \min_p \text{pen}(m, p, \delta, \gamma, n), \quad (12)$$

computable before observing the output values $\{y_i\}_{i=1}^n$, and then calculate the estimator $\hat{x}_{m,\hat{p}(m)} = R_{m,\hat{p}(m)}y$ which was defined in (2).

Finally, choose the best model as

$$\hat{m} = \arg \min_m L_{n,1}(y, \hat{x}_{m,\hat{p}(m)}, \hat{p}(m)). \quad (13)$$

The penalized estimator is then $\hat{x}_{\hat{m}} := \hat{x}_{\hat{m},\hat{p}(\hat{m})}$. It is important to remark that for each model m , $\hat{p}(m)$ is independent of y and hence of the random observation error structure. The following result assures the consistency of the proposed estimation procedure, although the obtained rates are not optimal as observed at the beginning of this section.

Theorem 2.5.1. *With probability greater than $1 - \delta$, we have*

$$\begin{aligned} L(\hat{x}_{\hat{m}}) &\leq \frac{1+\gamma}{1-4\gamma} [L(x_m) \\ &+ \min_{m,k} (2p_0(m, P_k, \delta) + \frac{1}{p_{\min}} \text{pen}_1(m, P_k, \delta) \\ &+ \frac{1}{p_{\min}^2} (1 + 2/\gamma) \text{pen}_2(m, P_k, \delta))] \\ &\leq \frac{1+\gamma}{1-4\gamma} \min_m [L(x_m) + \min_k \text{pen}(m, P_k, \delta, \gamma, n)] \end{aligned}$$

Proof. The proof follows from Lemma 2.5.2 below □

In order to state Lemma 2.5.2 we introduce for any given p and $x \in S_m, x' \in S_{m'}$, the quantities

$$\Delta_1(x, x', p) := [L_n(x, y, p) - L_n(x, p)] - [L_n(x', y, p) - L_n(x', p)]$$

and

$$\Delta_2(x, x', p) := [L_n(x, p) - L(x)] - [L_n(x', p) - L(x')].$$

We then have.

Lemma 2.5.2. *Let ε be a vector of i.i.d. random variables satisfying the moment condition [MC]. Assume that the conditions [AB], [AS] and [AQ] are satisfied. Assume that $p_k(t_i) > P_{k,\min}$ for all $i = 1, \dots, n$ and $L_{m,k} \geq 0$, such that the following Kraft inequality $\sum_{m,k} e^{-\sqrt{dr L_{m,k}(d_m+1)}} < 1$ holds. Assume pen_0 , pen_1 and pen_2 to be selected according to (9), (10) and (11) respectively. Let $x \in S_m$ and $x' \in S_{m'}$. Then*

$$\begin{aligned} &P(\sup_{m,m',k} (\Delta_1(x, x', P_k) \\ &- [\frac{1}{\gamma p_{\min}^2} (\text{pen}_2^2(m, P_k, \delta) + \text{pen}_2^2(m', P_k, \delta)) \\ &+ \gamma(\|x_0 - x\|_{n,q}^2 + \|x_0 - x'\|_{n,q}^2)]) > 0) \\ &\leq \delta/3 \end{aligned}$$

and

$$\begin{aligned}
& P\left(\sup_{m,m',k} \{\Delta_2(\hat{x}_{m,p}, x', P_k) \right. \\
& - [2p_0(m, P_k, \delta) + (\frac{1}{p_{\min}} + \frac{1}{\gamma})pen_1(m, P_k, \delta) \\
& + 2p_0(m', P_k, \delta) + \frac{1}{p_{\min}}pen_1(m', P_k, \delta) \\
& + 3\gamma\|x_0 - x_m\|_{n,q}^2 + (\frac{1}{p_{\min}^2}(\frac{1}{\gamma} + 1) + \frac{1}{\gamma})pen_2(m', P_k, \delta) \\
& \left. + 2((c+1)\frac{n^{-(1+\alpha)}QC}{p_{\min}})^2\} > 0\right) \\
& \leq 2\delta/3
\end{aligned}$$

Proof. In order to simplify notation throughout the proof of the lemma we use p instead of P_k .

The first part of Lemma 2.5.2 is rather standard, the only care being taking into account the random norm. For any $x \in S_m$ and $x' \in S_{m'}$ we have

$$\begin{aligned}
& |\Delta_1(x, x', p)| \\
& = \left| \frac{2}{n} \sum_{i=1}^n q_i \frac{w_i}{p_i} \varepsilon_i(x - x')(t_i) \right| \\
& \leq \left| \frac{2}{n} \sum_{i=1}^n q_i \frac{w_i}{p_i} \varepsilon_i(x_0 - x)(t_i) \right| \\
& \quad + \left| \frac{2}{n} \sum_{i=1}^n q_i \frac{w_i}{p_i} \varepsilon_i(x_0 - x')(t_i) \right| \\
& \leq 2\|(x_0 - x)\|_{n,qw/p} \sup_{\|v\|_{n,qw/p}=1, v \in S_m} \frac{1}{n} \sum_{i=1}^n q_i \frac{w_i}{p_i} \varepsilon_i v_i \\
& \quad + 2\|(x_0 - x')\|_{n,qw/p} \sup_{\|v\|_{n,qw/p}=1, v \in S_{m'}} \frac{1}{n} \sum_{i=1}^n q_i \frac{w_i}{p_i} \varepsilon_i v_i \\
& \leq \gamma(\|x_0 - x\|_{n,q}^2 + \|x_0 - x'\|_{n,q}^2) \\
& \quad + \frac{1}{\gamma p_{\min}^2} (\|R_{m,p}\varepsilon\|_{n,qw/p}^2 + \|R_{m',p}\varepsilon\|_{n,qw/p}^2),
\end{aligned}$$

where we have used $\|R_{m,p}\varepsilon\|_{n,qw/p} = \sup_{\|v\|_{n,qw/p}=1} \frac{1}{n} \sum_{i=1}^n q_i \frac{w_i}{p_i} \varepsilon_i v_i$ and the inequality $2ab \leq \gamma a^2 + 1/\gamma b^2$ to obtain the stated result.

Hence,

$$\begin{aligned}
& P_w\left(\sup_{m,m',p} [\Delta_1 \right. \\
& \quad - \frac{1}{p_{\min}^2 \gamma} (pen_2(m, p, \delta) + pen_2(m', p, \delta)) \\
& \quad \left. - \gamma(\|x_0 - x\|_{n,q}^2 + \|x_0 - x'\|_{n,q}^2)] > 0\right) \\
& \leq 2 \sum_{m,k} P_w(\|R_{m,p}\varepsilon\|_{n,qw/p}^2 > pen_2(m, p, \delta)) \\
& \leq \delta/3
\end{aligned} \tag{14}$$

by Lemma 2.4.4 and the choice of the penalization pen_2 in (11), recalling $R_{m,p}$ is a projection matrix.

The term Δ_2 requires a little more work. To begin with, for any $x \in S_m$, write $\widetilde{L}_n(x, p) := \|x - x_0\|_{n,qw/p}^2$ and $\widetilde{L}(x) := \|x - x_0\|_{n,q}^2$. Recall that, for a given m , $\hat{x}_{m,p} - x_0 = R_{m,p}(x_0 - x_m) + (x_m - x_0) + R_{m,p}\varepsilon$. To deal with this term, we must consider all the terms in the square of this expression. Thus,

$$\begin{aligned} \widetilde{L}_n(\hat{x}_{m,p}, p) - \widetilde{L}(\hat{x}_{m,p}, p) &= \frac{1}{n} \sum_{i=1}^n q(t_i) \left(\frac{w_i}{p_i} - 1 \right) [R_{m,p}(x_0 - x_m)]^2(t_i) \\ &+ \frac{2}{n} \sum_{i=1}^n q(t_i) \left(\frac{w_i}{p_i} - 1 \right) R_{m,p}(x_0 - x_m)(t_i) (x_m - x_0)(t_i) \\ &+ \frac{1}{n} \sum_{i=1}^n q(t_i) \left(\frac{w_i}{p_i} - 1 \right) (x_m - x_0)^2(t_i) \\ &+ \frac{1}{n} \sum_{i=1}^n q(t_i) \left(\frac{w_i}{p_i} - 1 \right) [R_{m,p}\varepsilon]_i^2 \\ &+ \frac{2}{n} \sum_{i=1}^n q(t_i) \left(\frac{w_i}{p_i} - 1 \right) [R_{m,p}\varepsilon]_i [R_{m,p}(x_0 - x_m)](t_i) \\ &+ \frac{2}{n} \sum_{i=1}^n q(t_i) \left(\frac{w_i}{p_i} - 1 \right) [R_{m,p}\varepsilon]_i (x_0 - x_m)(t_i) \\ &= I_a + I_b + I_c + I_d + I_e + I_f \end{aligned}$$

Start with I_a .

Write

$$\begin{aligned} &\|R_{m,p}(x_0 - x_m)(t_i)\|_{n,qw/p}^2 - \|R_{m,p}(x_0 - x_m)(t_i)\|_{n,q}^2 \\ &= \frac{1}{n} \sum_{i=1}^n q(t_i) \left(\frac{w_i}{p_i} - 1 \right) [R_{m,p}(x_0 - x_m)]^2(t_i). \end{aligned}$$

Note that

$$\begin{aligned} &\|[R_{m,p} - \mathbb{E}(R_{m,p})](x_0 - x_m)(t_i)\|_{n,qw/p}^2 - \|[R_{m,p} - \mathbb{E}(R_{m,p})](x_0 - x_m)(t_i)\|_{n,q}^2 \\ &\leq \frac{1}{p_{min}} \|[R_{m,p} - \mathbb{E}(R_{m,p})]\|_\rho \left(\|(x_0 - x_m)(t_i)\|_{n,q}^2 \right), \end{aligned}$$

Whence from the choice of pen_1 , using Lemma 2.4.1 and summing over m we obtain

$$\begin{aligned} &P(\sup_m \sup_p [\|[R_{m,p} - \mathbb{E}(R_{m,p})](x_0 - x_m)(t_i)\|_{n,qw/p}^2 \\ &- \|[R_{m,p} - \mathbb{E}(R_{m,p})](x_0 - x_m)(t_i)\|_{n,q}^2 - pen_1(m, P_k, \delta)]) \\ &< \delta/6. \end{aligned} \tag{15}$$

We then use Lemma 2.4.5 to bound $\|(\mathbb{E}(R_{m,p}) - R_m)[x_0 - x_m]\|_{n,q} = (c+1)n^{-1-\alpha}e_m/p_{min}$ and achieve the bound of the term I_a .

For the term I_b we start by remarking that

$$\sum_{i=1}^n q(t_i) \frac{w_i}{p_i} R_{m,p}(x_0 - x_m)(t_i)(x_m - x_0)(t_i) = 0$$

by orthogonality. The term

$$\sum_{i=1}^n q(t_i) R_{m,p}(x_0 - x_m)(t_i)(x_m - x_0)(t_i)$$

is then bounded

$$\left[\sum_{i=1}^n q(t_i) R_{m,p}(x_0 - x_m)(t_i)(x_m - x_0)(t_i) \right]^2 \leq \gamma \|x_0 - x_m\|_{n,q}^2 + \frac{1}{\gamma} \|R_{m,p}[x_0 - x_m]\|_{n,q}^2$$

and the proof follows as for I_a .

For I_c , the proof follows from Lemma 2.5.3 below,

$$P(\sup_m \sup_p \{ \frac{1}{n} \sum_i q(t_i) (\frac{w_i(p)}{p_i} - 1) (x_m - x_0)^2(t_i) - p_0(m, p, \delta) \} > 0) < \delta/6. \quad (16)$$

For I_d , Lemma 2.4.4 implies that

$$P(\sup_m \sup_p \{ \frac{1}{n} \sum_i q(t_i) (\frac{w_i(p)}{p_i} - 1) [R_{m,p}\varepsilon]_i^2 - \frac{1}{p_{\min}^2} p_2(m, p, \delta) \} > 0) < \delta/6. \quad (17)$$

The term I_e follows exactly as for Δ_1 . Finally, as for I_b , by orthogonality we only have to bound the term

$$\left[\sum_{i=1}^n q(t_i) [R_{m,p}\varepsilon]_i (x_m - x_0)(t_i) \right]^2$$

whose proof then follows exactly as for I_d .

The proof then follows by gathering the bounds in (14), (15), (16) and (17). \square

Lemma 2.5.3. *Assume that there exists a positive constant C such that $\sup_m e_m \leq C$ with $e_m = \|x_0 - x_m\|_\infty$. Assume that the condition $[AQ]$ is satisfied and that $p_k(t_i) > P_{k,\min}$ for all $i = 1, \dots, n$. Assume pen_0 to be selected according to (9). Then,*

$$P(\sup_m \sup_p \{ \|x_0 - x_m\|_{n,qw/p}^2 - \|x_0 - x_m\|_{n,q}^2 - pen_0(m, p, \delta) \} > 0) < \delta/6$$

Proof. Note that

$$\begin{aligned} & \|x_0 - x_m\|_{n,qw/p}^2 - \|x_0 - x_m\|_{n,q}^2 \\ &= \frac{1}{n} \sum_{i=1}^n q(t_i) \left(\frac{w_i}{p_i} - 1 \right) (x_0 - x_m)^2(t_i). \end{aligned}$$

Let p^* attain the supremum of this expression, so that

$$\begin{aligned} & \|x_0 - x_m\|_{n,qw/p^*}^2 - \|x_0 - x_m\|_{n,q}^2 \\ &= \sup_p \left\{ \|x_0 - x_m\|_{n,qw/p}^2 - \|x_0 - x_m\|_{n,q}^2 \right\}. \end{aligned}$$

Since $x - x_0$ is not random and is uniformly bounded

$$\begin{aligned} & \mathbb{E} \left(\sup_p \frac{1}{n} \sum_i q(t_i) \left(\frac{w_i(p)}{p_i} - 1 \right) (x_0 - x_m)^2(t_i) \right) \\ &= \frac{1}{n} \sum_i q(t_i) (x_0 - x_m)^2(t_i) \mathbb{E} \left(\left(\frac{w_i(p^*)}{p_i^*} - 1 \right) \right) = 0, \end{aligned}$$

Whence from the choice of pen_0 in (9), using the bounded differences inequality ([4]). Thus, we have

$$P(\sup_m \sup_p \{ \|x_0 - x_m\|_{n,qw/p}^2 - \|x_0 - x_m\|_{n,q}^2 - pen_0(m, p, \delta) \} > 0) \leq \sum_m \frac{\delta}{6m(m+1)},$$

which yields the desired result. \square

2.6. Error bounds for the general bounded case

The above procedure can be extended to other frameworks, defined by minimization over $S_0 = \bigcup S_m$ of a given loss function

$$L_n(x, y, p) = \frac{1}{n} \sum_{i=1}^n q_i \frac{w_i}{p_i} l(x(t_i), y_i)$$

with expectation $L(x) = \mathbb{E}(L_n(x, y, p)) = \frac{1}{n} \sum_{i=1}^n q_i \mathbb{E}(l(x(t_i), y(t_i)))$. Set, as above, $L_n(x, p) = \mathbb{E}_\varepsilon(L_n(x, y, p)) = \frac{1}{n} \sum_{i=1}^n q_i \frac{w_i}{p_i} \mathbb{E}(l(x(t_i), y(t_i)))$. We will denote $l(x) = \mathbb{E}(l(x, y))$. In order to repeat the proof of section 2.5 it is necessary to control both the fluctuations of $L_n(x, y, p) - L_n(x', y, p) - [L_n(x, p) - L_n(x', p)]$ and $L_n(x, p) - L_n(x', p) - [L(x) - L(x')]$. The first term typically requires setting bounds for

$$\Delta(m, p) := \sup_{x \in S_m} \frac{1}{n} \sum_{i=1}^n q_i \frac{w_i}{p_i} [l(x(t_i), y_i) - l(x(t_i))].$$

Assuming $l(x, y)$ is uniformly bounded by a constant B^2 (which is not the case for the example presented in section 2.5) standard arguments (see for example [4] for a very thorough discussion), combining the bounded differences inequality and bounds for Radamacher sums lead to

$$P(\Delta(m, P_k) - [t_{m,k} + \mathbb{E}(2R_n(m, k))] > 0) \leq \frac{\delta}{k(k+1)m(m+1)},$$

with

$$t_{m,k} = \sqrt{\frac{4B^4 \log(2m(m+1)k(k+1)/\delta)}{np_{\min}^2}}$$

and

$$R_n(m, k) = \mathbb{E}_\sigma \left(\sup_{x \in S_m} \frac{1}{n} \sum_{i=1}^n q_i \frac{w_i}{p_i} \sigma_i l(x(t_i), y_i) \right)$$

where σ_i is a sequence of independent (Radamacher) random variables, $P(\sigma = -1) = P(\sigma = 1) = 1/2$ and independent of y_i . The above discussion yields,

$$P(\sup_{m,k} \Delta(m, P_k) - [t_{m,k} + 2\mathbb{E}(R_n(m, k))] > 0) < \delta \quad (18)$$

For the second term it is then necessary to bound

$$\Delta'(m, k) := \sup_{x \in S_m} \frac{1}{n} \sum_{i=1}^n q_i \left(\frac{w_i}{p_i} - 1 \right) l(x(t_i)).$$

If $l(x)$ is bounded, again combining the bounded differences inequality and bounds for Radamacher sums lead to

$$P(\Delta'(m, P_k) - [t_{m,k} + 2\mathbb{E}(R'_n(m, k))] > 0) \leq \frac{\delta}{k(k+1)m(m+1)},$$

with

$$t_{m,k} = \sqrt{\frac{4B^4 \log(2m(m+1)k(k+1)/\delta)}{np_{\min}^2}}$$

and

$$R'_n(m, k) = \mathbb{E}_\sigma \left(\sup_{x \in S_m} \frac{1}{n} \sum_{i=1}^n q_i \frac{w_i}{p_i} \sigma_i l(x(t_i)) \right).$$

This yields,

$$P(\sup_{m,k} \Delta'(m, P_k) - [t_{m,k} + 2\mathbb{E}(R'_n(m, k))] > 0) < \delta \quad (19)$$

Actually, the bounds in section 2.5 follow from bounding the Radamacher sums in that case by $R_n(m, k) \leq CTr(R_{n, P_k})$, for a certain constant C , using Lemma 2.4.4 (which follows from a functional exponential inequality proved in [5]) and choosing adequate terms $L_{k,m}$ in order to assure converge of the sum. Hence, it would seem that equation (19) does not add any interesting information to what has already been discussed. However, the more general setting is important because (in the bounded case) allows us to pass from Radamacher sums to V-C dimensions (see again [4] for a general discussion) which allows us in turn to consider more general solution spaces (than a numerable union of target model spaces S_m and a numerable collection of target probabilities). Bounds in this case would be

$$P(\sup_{x \in S_{0,p}} [L_n(x, y, p) - L(x)] > 2[\sqrt{\frac{4B^4 \log(2/\delta)}{np_{\min}^2}} + 2\sqrt{\frac{2 \log(n)V}{np_{\min}}}] < \delta \quad (20)$$

where V is the V-C dimension of the class of functions S_0 .

In practice a reasonable alternative is estimating the overall error by cross-validation or leave one out techniques and then choose m minimizing the error for successive essays of probability \hat{p} . Recall, in the procedure of section 2.5, labels are not required to obtain \hat{p} . Of course this requires a stock of “extra” labels, which might not be affordable in the active learning setting. However, many applications suggest that \hat{p} (or a threshold version of \hat{p} which eliminates points with sampling probability $\hat{p}_i \leq \eta$ a certain small constant) helps finding “good” or informative subsets, over which model selection may be performed. Other intermediate versions of error bounding include using empirical versions of the V-C dimension. Empirical error minimization is specially useful for applications where what is required is a subset of very informative sample points, as for example when deciding what points get extra labels (new laboratory runs, for example) given a first set of complete labels is available.

3. Iterative procedure: updating the sampling probabilities

A major drawback of the batch procedure is the appearance of p_{\min} in the denominator of error bounds, since typically p_{\min} must be small in order for the estimation procedure to be effective. Indeed, since the expected number of effective samples is given by $n_s := \mathbb{E}(\sum_i p_i)$, small values of p_i are required in order to gain in sample efficiency.

A closer look at the proofs shows it is necessary to improve on the bounds of expressions such as

$$\frac{1}{n} \sum_{i=1}^n q_i \frac{w_i}{p_i} \varepsilon_i(x - x')(t_i)$$

or

$$\frac{1}{n} \sum_{i=1}^n q_i \left(\frac{w_i}{p_i} - 1 \right) (x - x')^2(t_i).$$

Thus, it seems like a reasonable alternative to consider iterative procedures for which at time j $p_j(i) \sim \max_{x, x' \in S_j} |x(t_i) - x'(t_i)|$ with S_j the current hypothesis space. In what follows we develop this strategy, adapting the results of [2] from the classification to the regression problem. Although we continue to work in the setting of model selection over bounded subsets of linearly spanned spaces, results can be readily extended to other frameworks such as additive models or kernel models. Once again, we will require certain additional restrictions associated to the uniform approximation of x_0 over the target model space.

More precisely. We start with an initial model set $S(= S_{m_0})$ and set x^* to be the overall minimizer of the loss function $L(x)$ over S . Assume additionally

$$\mathbf{AU} \quad \sup_{x \in S} \max_{t \in \{t_1, \dots, t_n\}} |x_0(t) - x(t)| \leq B$$

Let $L_n(x) = L_n(x, y, p)$ and $L(x)$ be as in section 2.5. For the iterative procedure introduce the notation

$$L_j(x) := \frac{1}{n_0 + j} \sum_{\ell=1}^{n_0+j} q(t_{\ell_j}) \frac{w_j}{p_j} (x(t_{\ell_j}) - y(t_{\ell_j}))^2.$$

In the setting of Section 2 for each $0 \leq j \leq n$, S_j will be the linear space spanned by the collection $\{\phi_\ell\}_{\ell \in I_j}$ with $|I_j| = d_j$.

In order to bound the fluctuations of the initial step in the iterative procedure we consider the quantities defined in equations (5) and (7) for $r = \gamma = 2$. That is,

$$\begin{aligned} \Delta_0 &= 2\sigma^2 Q \left\{ \frac{2(d_0 + 1)}{n_0} + \frac{\log^2(2/\delta)}{n_0} \right\} \\ &\quad + 2(\tilde{\beta}_{m_0}(1 + \tilde{\beta}_{m_0}))^2 B^2. \end{aligned}$$

with

$$\tilde{\beta}_{m_0} = \frac{c_{m_0}(\sqrt{17} + 1)}{2} \sqrt{\frac{d_0 Q}{n_0 p_{\min}}} \sqrt{2 \log(2^{7/4} m_0 / \delta)}.$$

As discussed in section 2.4, Δ_0 requires some initial guess of $\|x_0 - x_{m_0}\|_{n,q}^2$. Since this is not available, we consider the upper bound B^2 . Of course this will possibly slow down the initial convergence as Δ_0 might be too big, but will not affect the overall algorithm. Also remark we do not consider the weighting sequence L_k of equation (7) because the sampling probability is assumed fixed.

Next set $B_j = \sup_{x, x' \in S_{j-1}} \max_{t \in \{t_1, \dots, t_n\}} |x(t) - x'(t)|$ and define

$$\begin{aligned} \Delta_j = & \sqrt{\sigma^2 Q \left[\left(\frac{2(d_j + 1)}{n_0 + j} \right) + \frac{\log^2(4(n_0 + j)(n_0 + j + 1)/\delta)}{n_0 + j} \right]} \\ & + \sqrt{\log(4(n_0 + j)(n_0 + j + 1)/\delta) \frac{16B_j^2(2B_j \wedge 1)^2 Q^2}{n_0 + j}} + 4\sqrt{4 \frac{(d_j + 1) \log n}{n_0 + j}}. \end{aligned}$$

The iterative procedure is stated as follows:

1. For $j = 0$:
 - Choose (randomly) an initial sample of size n_0 , $M_0 = \{t_{i_1}, \dots, t_{i_{n_0}}\}$.
 - Let \hat{x}_0 be the chosen solution by minimization of $L_0(x)$ (or possibly a weighted version of this loss function).
 - Set $S_0 \subset \{x \in S : L_0(x) < L_0(\hat{x}_0) + \Delta_0\}$
2. At step j :
 - Select (randomly) a sample candidate point t_{i_j} , $t_{i_j} \notin M_{j-1}$. Set $M_j = M_{j-1} \cup \{t_{i_j}\}$
 - Set $p_j = (\max_{x, x' \in S_{j-1}} |x_{t_j} - x'_{t_j}| \wedge 1)$ and generate $w_j \sim \text{Ber}(p_j)$. If $w_j = 0$, set $j = j + 1$ and go to (2) to choose a new sample candidate. If $w_j = 1$ sample y_{t_j} and continue.
 - Let $\hat{x}_j = \arg \min_{x \in S_{j-1}} L_j(x) + \Delta_{j-1}(x)$
 - Set $S_j \subset \{x \in S_{j-1} : L_j(x) < L_j(\hat{x}_j) + \Delta_j\}$
 - Set $j = j + 1$ and go to (2) to choose a new sample candidate.

Remark, that such as it is stated, the procedure can continue only up until time n (when there are no more points to sample). If the process is stopped at time $T < n$ the term $\log(n(n+1))$ can be replaced by $\log(T(T+1))$.

Also, instead of the term $4\sqrt{4 \frac{(d_j+1) \log n}{n_0+j}}$ in the second line of the definition of Δ_j , we could have used $4R_j$, the associated Radamacher sum. Recall the Radamacher sum over the class S_{j-1} is given by

$$R_j = \mathbb{E}_\sigma \left(\sup_{x \in S_{j-1}} \frac{1}{j + n_0} \sum_i \sigma_i q_i \frac{w_i}{p_i} (x - x_0)^2(t_{\ell_i}) \right).$$

As discussed in section 2.6, $R_j \leq R_0 \leq \sqrt{\frac{2 \log(n) V_S}{n}}$, where V_S is the V-C dimension of the class of functions $\mathcal{F} = \{f = (x - x_0)^2, x \in S\}$. The quantity $2(d_j + 1)$ in the definition of Δ_j is obtained by using well known properties of the V-C dimension, using x^2 is a convex function and $d_j + 1$ is the V-C dimension of the linear space $S'_j = S_j - x_0$, to obtain the stated weight.

We have the following result, in the spirit of Theorem 2 in [2].

Theorem 3.0.1. *Let $x^* = \arg \min_{x \in S} L(x)$. Set $\delta > 0$. Then, with probability at least $1 - \delta$ for any $j \leq n$*

- $|L(x) - L(x')| \leq 2\Delta_{j-1}$, for all $x, x' \in S_j$
- $L(\hat{x}_j) \leq [L(x^*) + 2\Delta_{j-1}]$

An important issue is related to the initial choice of m_0 and n_0 . As the overall precision of the algorithm is determined by $L(x^*)$, it is important to select a sufficiently complex initial model collection. However, if $d_{m_0} \gg n_0$ then Δ_0 can be big and $p_j \sim 1$ for the first samples, which leads to a more inefficient sampling scheme.

Proof of Theorem 3.0.1: the proof is based on the following preliminary Lemma.

Lemma 3.0.2. *For any $\delta > 0$, with probability at least $1 - \delta$, for all $j \leq n$ and all $x, x' \in S_{j-1}$*

$$|L_j(x) - L_j(x') - [L(x) - L(x')]| \leq \Delta_j.$$

Set $\delta > 0$ so from lemma 3.0.2

$$|L_j(x) - L_j(x') - [L(x) - L(x')]| \leq \Delta_j$$

holds for all $0 \leq j \leq n$, $x, x' \in S_{(j-1) \vee 0}$ with probability at least $1 - \delta$. Hence, for any $x, x' \in S_{j-1}$, since $S_{j-1} \subset S_{j-2}$, from the definition of S_{j-2}

$$\begin{aligned} L(x) - L(x') &\leq L_{j-1}(x) - L_{j-1}(x') + \Delta_{j-1} \\ &\leq L_{j-1}(\hat{x}_{j-1}) + \Delta_{j-1} - L_{j-1}(\hat{x}_{j-1}) + \Delta_{j-1} = 2\Delta_{j-1}. \end{aligned}$$

On the other hand, over the chosen event with probability greater than $1 - \delta$, by the choice of Δ_0 and the results in section 2.4, $x^* \in S_0$ from the definition of S_0 . We shall now prove by induction that over the stated event $x^* \in S_j$ for $1 \leq j \leq n$. Assume $x^* \in S_{j-2}$. By lemma 3.0.2,

$$L_{j-1}(x^*) - L_{j-1}(\hat{x}_{j-1}) \leq L(x^*) - L(\hat{x}_{j-1}) + \Delta_{j-1} \leq \Delta_{j-1},$$

so that $x^* \in S_{j-1}$, which ends the proof by induction. Whence, for all $1 \leq j \leq n$, $L(\hat{x}_j) \leq L(x^*) + 2\Delta_{j-1}$, which ends the proof of the Theorem.

Proof of lemma 3.0.2:

For fixed j and any $x, x' \in S_{j-1}$ we have

$$\begin{aligned} &L_j(x) - L_j(x') - [L(x) - L(x')] \\ &= \frac{1}{j + n_0} \sum_{i=1}^{n_0+j} q_i \left(\frac{w_i}{p_i} - 1 \right) (x - x')(t_{\ell_i})(x + x' - 2x_0)(t_{\ell_i}) \\ &\quad + \frac{2}{j + n_0} \sum_{i=1}^{n_0+j} q_i \frac{w_i}{p_i} \varepsilon_i(x(t_{\ell_i}) - x'(t_{\ell_i})) \\ &= I_j + II_j. \end{aligned}$$

Set $z_i = q_i w_i (x - x')(t_{\ell_i})(x + x' - 2x_0)(t_{\ell_i})$, so that $\|z_i\|_\infty \leq 2QB_j(2B_j \wedge 1)$ and I_j satisfies the bounded difference inequality with $c^2 = 4Q^2B_j^2(2B_j \wedge 1)^2$. By equation (19) in section 2.6 we have

$$\begin{aligned} &P(I_j > \sqrt{\log(4(n_0 + j)(n_0 + j + 1)/\delta)} \frac{16B_j^2(2B_j \wedge 1)^2Q^2}{n_0 + j} + 4\sqrt{4 \frac{(d_j + 1) \log n}{n_0 + j}}) \\ &\leq P(I_j > \sqrt{\log(4(n_0 + j)(n_0 + j + 1)/\delta)} \frac{16B_j^2(2B_j \wedge 1)^2Q^2}{n_0 + j} + 4\mathbb{E}(R_j)) \\ &\leq \delta/2((n_0 + j)(n_0 + j + 1)). \end{aligned}$$

Next we deal with II_j . Set $u(t) = \frac{w_i(x-x')(t)}{p_i} \frac{1}{\sqrt{Q}}$, so that $\frac{1}{n_0+j} \sum_i q_i u^2(t_{\ell_i}) \leq 1$. Then,

$$\begin{aligned} |II_j| &\leq \sup_{u \in S_{j-1}, \|u\|_{n_0+j,q}} \frac{\sqrt{Q}}{n_0 + j} \sum_i q_i u_i \varepsilon_i \\ &= \sqrt{Q} \|\Pi_{S_{j-1}} \varepsilon\|_{n_0+j,q}, \end{aligned}$$

where Π_{S_j} stands for the projection over S_j . Whence by Lemma 2.4.4,

$$\begin{aligned} P \left(II_j > \sqrt{\sigma^2 Q \left[\left(\frac{2(d_j + 1)}{n_0 + j} \right) + \frac{\log^2(4(n_0 + j)(n_0 + j + 1)/\delta)}{n_0 + j} \right]} \right) \\ \leq \delta/2((n_0 + j)(n_0 + j + 1)). \end{aligned}$$

Summing over j ends the proof.

3.1. Effective sample size

For any sampling scheme the expected number of effective samples is, as already mentioned, $\mathbb{E}(\sum_i p_i)$. Whenever the sampling policy is fixed, this sum is not random and effective reduction of the sample size will depend on how small sampling probabilities are. However, this will increase the error bounds as a consequence of the factor $1/p_{\min}$. The iterative procedure allows a closer control of both aspects and under suitable conditions will be of order $\sum_j \sqrt{(L(x^*) + \Delta_j)}$, as we will prove below establishing appropriate bounds over the random sequence p_j . Recall from the definition of the iterative procedure we have $p_j(i) \sim \max_{x, x' \in S_j} |x(t_i) - x'(t_i)|$, whence the expected number of effective samples is of the order of $\sum_j \max_{x, x' \in S_j} |x(t_i) - x'(t_i)|$. It is then necessary to control $\sup_{x, x' \in S_{j-1}} |(x - x')(t_i)|$ in terms of the (quadratic) empirical loss function L_j . For this we must introduce some notation and results relating the supremum and L_2 norms ([3]).

Let $S \subset L_2 \cap L_\infty$ be a linear subspace of dimension d , with basis $\Phi := \{\phi_j, j \in m_S\}$, $|m_S| = d$. Set $\eta(S) := \frac{1}{\sqrt{d}} \sup_{x \in S, x \neq 0} \frac{\|x\|_\infty}{\|x\|_2}$, $r_\Phi := \frac{1}{\sqrt{d}} \sup_{\beta, \beta \neq 0} \frac{\sum_{\phi_j \in \Phi} \beta_j \phi_j}{\|\beta\|_\infty}$ and $\bar{r} := \inf_\Lambda r_\Lambda$, where Λ stands for any orthonormal basis of S . We require the following result in [3]:

Lemma 3.1.1. (Lemma 1 [3]) *Let S be a d dimensional linear subspace of $L_2 \cap L_\infty$, with basis Φ , and set $\eta(S) := \frac{1}{\sqrt{d}} \sup_{t \in S} \|t\|_\infty / \|t\|_2$. Then*

1. $\eta(S) = \|\sum_{\phi_j \in \Phi} \phi_j^2\|_\infty / 2$
2. $\eta(S) \leq \bar{r} \leq \eta(S) \sqrt{d}$

Example 3.1.2. *Some examples of \bar{r} for typical linear settings include ([3], pp 337–338):*

1. *Trigonometric expansions:* $\bar{r} \leq \sqrt{2d}$.
2. *Polynomials:* $\bar{r} \leq d$.
3. *Localized basis:*
 - $\{\phi_j = \sqrt{d} 1_{[(j-1)/d, j/d]}\}_{1 \leq j \leq d}$: $\bar{r} \leq 1$
 - *Piecewise polynomials on $[0, 1]$ of degree m :* $\bar{r} \leq 2m + 1$
 - *Orthonormal wavelet systems:* $\bar{r} \leq C$, for a certain constant C depending on the form of the basis.

We have the following result

Lemma 3.1.3. *Let \hat{x}_j be the sequence of iterative approximations to x^* and $p_t(j)$ be the sampling probabilities in each step of the iteration, $j = 1, \dots, T$. Then, the effective number of samples, that is, the expectation of the required samples $n_s = \mathbb{E} \left(\sum_{j=1}^T p_j(t_j) \right)$ is bounded by*

$$n_s \leq 2\sqrt{2}\bar{r}(\sqrt{L(x^*)} \sum_{j=1}^T \sqrt{d_j} + \sum_{j=1}^T \sqrt{d_j \Delta_j}).$$

Proof. We have

$$\begin{aligned}
(p_t(j))^2 &\leq \sup_{x, x' \in S_j} \|x - x'\|_\infty^2 \leq 4 \sup_{x \in S_j} \|x - x^*\|_\infty^2 \\
&\leq 4\bar{r}^2 d_j \sup_{x \in S_j} L(x - x^*) \\
&\leq 4\bar{r}^2 d_j \sup_{x \in S_j} [L(x) + L(x^*)] \\
&\leq 4\bar{r}^2 d_j (2L(x^*) + 2\Delta_j).
\end{aligned}$$

The third inequality follows from Lemma 3.1.1 and the fifth from the bound $\sup_{x \in S_j} L(x) \leq L(x^*) + 2\Delta_j$ as follows from Theorem 3.0.1. The proof is achieved by calculating the square root of each side of the last series of inequalities and finally using that $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$. \square

4. Appendix

Proof. (Lemma 2.4.2)

The proof follows closely the ideas of the proof of Theorem 7.3 p. 62 in [8]. Let $A \in \mathbb{R}^{n \times m}$ be a matrix, with rows $a(l) \in \mathbb{R}^m$, $l=1, \dots, n$, satisfying

$$\|a(l)\|_2 \leq K\sqrt{m} \quad (21)$$

for some constant $K \geq 1$. Recall $A^T A = \sum_{l=1}^n a(l)a(l)^T$.

For the first part of the lemma we must bound $E_r = \mathbb{E} \left(\left\| \frac{1}{n\Lambda_A} (A^T A - \mathbb{E}(A^T A)) \right\|_\rho^r \right)$, where $\Lambda_A := \frac{1}{n} \|\mathbb{E}(A^T A)\|_\rho$. Using the symmetrization Lemma (see [8]), we have for all $2 \leq r < \infty$,

$$E_r \leq \left(\frac{2}{n\Lambda_A} \right)^r \mathbb{E} \left(\|\epsilon_l a(l)a(l)^T\|_\rho^r \right).$$

where $\epsilon = (\epsilon_1, \dots, \epsilon_n)$ is a Rademacher sequence independent of $a(1), \dots, a(n)$.

Thus, the following inequality holds

$$E_r \leq \left(\frac{2}{n\Lambda_A} \right)^r 2^{3/4} m r^{r/2} e^{-r/2} \mathbb{E} \left(\|A\|_{\rho, l=1, \dots, n}^r \max_{l=1, \dots, n} \|a(l)\|_2^r \right),$$

where we have used Rauhut's Lemma 6.18, p. 46 [8] (which is a version of Rudelson's Lemma [9]) and the Cauchy-Schwarz inequality to obtain the stated result.

Furthermore, using the bound (21) and applying the triangle inequality yields

$$\begin{aligned}
&\mathbb{E} \left(\|A\|_{\rho, l=1, \dots, n}^r \max_{l=1, \dots, n} \|a(l)\|_2^r \right) \\
&\leq \sqrt{\mathbb{E} (\|A^T A\|_\rho^r) \mathbb{E} \left(\max_{l=1, \dots, n} \|a(l)\|_2^{2r} \right)} \\
&\leq K^r m^{r/2} (n\Lambda_A)^{r/2} ((\mathbb{E} \left(\left\| \frac{1}{n\Lambda_A} (A^T A - \mathbb{E}(A^T A)) \right\|_\rho^r \right))^{1/r} + 1)^{r/2}, \quad (22)
\end{aligned}$$

where we have used $\|\mathbb{E}(1/n A^T A) / \Lambda_A\|_\rho = 1$. Now, recall

$$\Sigma_{r,m,n} = \left(\frac{2K}{\sqrt{n\Lambda_A}} \sqrt{\frac{m}{n}} \right)^r 2^{3/4} m r^{r/2} e^{-r/2}.$$

Then, using the inequality (22)

$$E_r \leq \Sigma_{r,m,n}(E_r^{1/r} + 1)^{r/2}.$$

Whence, squaring the last inequality and completing squares yields

$$\left(E_r^{1/r} - \Sigma_{r,m,n}/2\right)^2 \leq \Sigma_{r,m,n}^2 + \Sigma_{r,m,n}^4/4.$$

In the following we assume that $\Sigma_{r,m,n} \leq 1/2$. Thus,

$$E_r^{1/r} \leq \sqrt{17}/4 \Sigma_{r,m,n} + 1/4 \Sigma_{r,m,n} = \tau \Sigma_{r,m,n}.$$

where $\tau = (\sqrt{17} + 1)/4$.

For the second part of the lemma we want to bound in probability $\|\frac{1}{n\Lambda_A}(A^T A - \mathbb{E}(A^t A))\|_\rho$. The proof then follows directly from the first part of the lemma using the Markov inequality. \square

References

- [1] Y. Barraud. Model selection for regression on a fixed design. *Probab. Theory Relat. Fields*, 117:467–493, 2000.
- [2] A. Beygelzimer, S. Dasgupta and J. Langford. Importance weighted active learning. *Proceedings of the 26th Annual International Conference on Machine Learning*, 2009.
- [3] L. Birgé, and P. Massart. Minimum Contrast Estimators on Sieves: Exponential Bounds and Rates of Convergence. *Bernoulli*, 4:, 329–395, 1998.
- [4] S. Boucheron, O. Bousquet and G. Lugosi. Theory of Classification : A survey of some recent advances. *ESAIM: Probability and Statistics*
- [5] O. Bousquet. Concentration Inequalities for Sub-Additive Functions Using of Entropy Method. *Stochastic Inequalities and Applications*, 56:213–247, 2005.
- [6] S. Hanneke. A bound on the label complexity of agnostic active learning. *Proceedings of the 24th International Conference on Machine Learning*, pp. 353–360, 2007.
- [7] A. Fermin and C. Ludeña. A statistical view of iterative methods for linear inverse problems. *TEST*, 17(2):381–400, 2008.
- [8] H. Rauhut. Compressive sensing and structured random matrices. *Radon Series Comp. Appl. Math* XX, 1–94, 2011.
- [9] M. Rudelson, Random vectors in the isotropic position. *J. Funct. Anal.*, 164:60–72, 1999.
- [10] M. Sugiyama and N. Rubens. A batch ensemble approach to active learning with model selection. *Neural Networks*, 21(9):1278–1286, 2008.